A Peer-To-Peer Agent-Based Semantic Search Engine

D. Beneventano[1,2], S. Bergamaschi[1,2], A. Fergnani[1], F. Guerra[1], M. Vincini[1], and D. Montanari[3,1]

[1] Dipartimento di Ingegneria dell'Informazione
Università di Modena e Reggio Emilia
Via Vignolese 905, 41100 Modena, Italy
{domenico.beneventano,sonia.bergamaschi,francesco.guerra,maurizio.vincini}@unimo.it
[2] IEIIT-BO Bologna - V.le Risorgimento 2, 40136 Bologna, Italy
[3] EniData, Viale Aldo Moro 38, 40127 Bologna, Italy
Daniele.Montanari@enidata.it

Abstract. Several architectures, protocols, languages, and candidate standards, have been proposed to let the "semantic web" idea take off. In particular, searching for information requires cooperation of the information providers and seekers. Past experience and history show that a successful architecture must support ease of adoption and deployment by a wide and heterogeneous population, a flexible policy to establish an acceptable cost-benefit ratio for using the system, and the growth of a cooperative distributed infrastructure with no central control.
In this paper an agent-based peer-to-peer system architecture to support search for information through a flexible integration of semantic information is defined. Two levels of integration are foreseen: strong integration of sources related to the same domain into a single information node by means of a mediator-based system; weak integration of information nodes on the basis of semantic relationships existing among concepts of different nodes. The system architecture is an evolution of the EU IST SEWASIE project. SEWASIE (http://www.sewasie.org/) aims at implementing an advanced search engine, which will provide SMEs with intelligent access to heterogeneous information on the Internet.

1 Introduction

Peer-to-peer (hereafter P2P) consists of an open-ended network of distributed computational *peers*, where each peer can exchange data and services with a set of other peers called *acquaintances*. Peers should be autonomous in choosing their

acquaintances. Moreover, it is usually assumed that there is no global control in the form of a global registry, global services, or global resources management nor a global schema or data repository. Gnutella and Napster [16] made the P2P paradigm popular as a version of distributed computing between traditional distributed systems and the web. Very recently a proposal in data management raised by this paradigm has been presented in [6]. In this context, each peer may have data to share with other peers and, in [6], it is assumed that each peer's database is relational and, since the data residing in different databases may have semantic inter-dependencies, peers are allowed to specify *coordination formulas*. Coordination formulas explain how data in one peer must relate to data in an acquaintance and may also act as constraints or for propagating updates. Peer' s need an acquaintance initialization protocol where two peers exchange views of their databases and agree on levels of coordination and the level of coordination should be dynamic, i.e. peers should be able to establish and modify acquaintances, with little human intervention. This is a crucial point and introduces a high degree of innovation into the traditional distributed databases and multi-database systems data management approach. The common assumption in this area is, in fact, to have a global database schema, usually obtained by skilled databases designers [21, 17]. In the new dynamic setting of P2P, we cannot assume the existence of a global schema for all databases in a P2P network or even those of the acquainted databases. Nevertheless, as proposed both in [6, 19] and in this paper, the architecture of heterogeneous distributed databases or often called *multi-database systems*, e.g. Multibase [18] , MOMIS [4, 2, 5], Garlic [8], TSIMMIS [9], and Information Manifold [13] is still valid. In most of these systems, a user issues queries to a global schema, and the system (called a *mediator* in [20]) maps the queries to subqueries on the underlying data sources. Each data source has a wrapper able to map subqueries into its native query language. A database designer is responsible for creating the global schema and the mappings with the data sources and for maintaining the schema and mappings with respect to evolution (i. e. data sources entering and leaving the system). The global schema so far obtained represents an ontology and a semantic enrichment of the underlying data sources.

SEWASIE system organizes and manages information in SINodes which follow the architecture of heterogeneous multi-database systems. In this paper we want to propose an evolution of the architecture of the SEWASIE system in the direction of the P2P which shares ideas with [19] and the JXTA search architecture [12]. In particular, we agree with [19] that it is no longer realistic to assume that the involved data sources act as if they were a single (virtual)

source, modeled as a global schema, as is done in classical data integration approaches. We propose an approach where we add to the role of a single virtual global schema a P2P architecture relying on a limited shared (or: overlapping) vocabularies between peers. Since overlaps between vocabularies of peers will be limited, query processing will have to be approximate. The result is a flexible architecture for query-processing in large, distributed and heterogeneous environments, based on a formal foundation. Further, we follow the JXTA search architecture [12] where the network of node peers holds at two levels: deep (i.e. data sources within an SINode) and wide (i.e. inter-SINodes).

After a presentation of the main features of the project, section 2, a proposal of architectural evolutions of the SEWASIE system in the direction of peer-to-peer paradigm is proposed in section 3.

2 The SEWASIE project

SEWASIE (SEmantic Webs and AgentS in Integrated Economies) (IST-2001-34825) is a research project founded by EU on action line Semantic Web (May 2002/April 2005) (http://www.sewasie.org/). The goal of the SEWASIE project is to design and implement an advanced search engine enabling intelligent access to heterogeneous data sources on the web via semantic enrichment to provide the basis of structured secure web-based communication. A SEWASIE user has at his disposal a search client with an easy-to-use query interface able to extract the required information from the Internet and to show it in an easily enjoyable format. From an architectural point of view, the SEWASIE prototype will provide a search engine client and indexing servers and ontologies.

The project will develop an agent-based secure, scalable and distributed system architecture for semantic search using community-specific multilingual ontologies. The developed system have to meet the needs of SMEs in a EU context.

The SEWASIE vision helps European enterprises to compete in a global market and to form strategic alliances by providing a sophisticated retrieval, brokering and communication service on basis of the semantic web technology. In particular, SEWASIE aims to help European SMEs to find strategic information, to provide advanced and novel services for monitoring and linking information in the context of risk management and competitor analysis, to provide ontology-based communication mechanisms for negotiation in multi-language environments.

3 SEWASIE in a P2P Architecture

3.1 Global architecture

The first basic idea underlying the architecture is that *to scale any Internet-based system well we need to decentralise control in a way resembling the decentralised control of the overall Internet architecture itself.* Currently available search engines are single, central entities (although they may actually be based on hundreds or thousands of machines) [7]. They need to concentrate metainformation on sources at a central location, and the user queries need to be processed at a central location as well. As a consequence, to prevent the collapse of the system under the increasing load of requests or storage requirements, the processing of the requests has to remain as simple as possible and the acquired metainformation must give up most or all the associated semantic. If a distributed architecture is used, then the number of processing points may scale and support increasing numbers of information nodes and information seekers, both in the metainformation acquisition phase and in the user request processing phase.

The second idea is that *semantic enrichment of data sources is the next step towards building information systems that are really useful.* However, the addition of semantics to data sources is a formidable task and it may be achieved only if info seekers and info providers may reach each other across a middle ground. This requires a *common* language and strategy, and the tools that actually flesh them both out.

The third idea is that *we have to deal with two level of knowledge.* We should envision a multi-level architecture, with local nodes and communities with strong ties as to develop a strong integration of their knowledge and information, into a global integrated ontology, while at a wider level the relationships among distinct nodes are established by means of weaker semantics mappings. The latter are maintained by an infrastructure of *brokers*, which will provide the entry points to the system and some routing of the queries towards the relevant information nodes.

A search system architecture satisfying the aforementioned ideas and desiderata is described in figure 1.

The **information nodes** (SINodes) groups together the modules which work to define and maintain a global integrated ontology presented to the network. A single information node may comprise several different systems.

The **brokering agents (BAs)** are the peers responsible for maintaining a view of the knowledge handled by the network, as well as the information on the specific content of SINodes which are under direct control (of each brokering

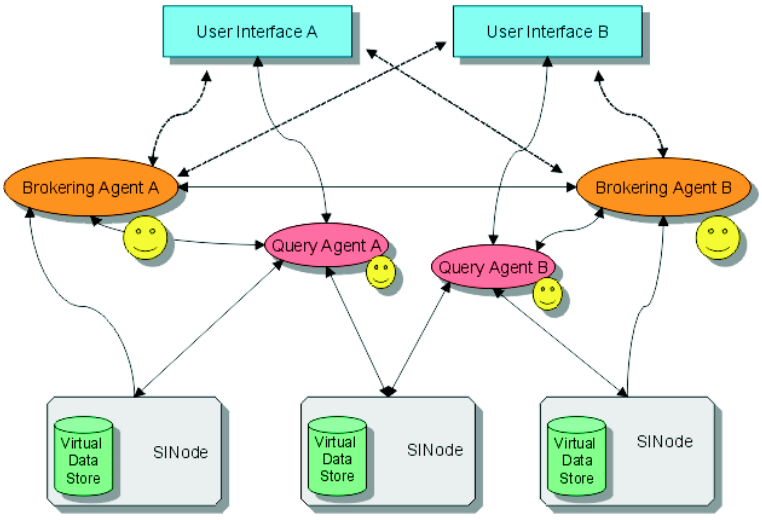

Fig. 1. General Overview of the peer-to-peer SEWASIE architecture

agent). These agents are intermediaries which have direct control over a number of SINodes, and provide the means to publish a manifesto within the network of the locally held information with a semantic profile.

The **query agents (QAs)** are the carriers of the user query from the user interface to the SINodes, and have the task of solving a query by interacting with the brokering agent network. Starting from a user- or task- specified brokering agent, they may access other BAs, connect with other information nodes, collect partial answers, and integrate them.

The **user interface** is the group of modules which work together to offer an integrated user interaction with the semantic search system. This interface needs to be personalized and configured with the specific user profile and a reference to the ontologies which are commonly used by this user.

The deployment of this architecture may achieve one of several different configuration:

- the *degree of expansion of the system*, which may go from "narrow" (limited diffusion, central control) up to "wide" (wide diffusion, distributed control)
- the *degree of intrinsic integrability* of different ontologies in the system, which may go from "deep" (great potential for integrability) to "shallow" (little or nothing is guaranteed)

The "narrow-deep" scenario is envisioned as typical of relatively contained, strongly organized environments. An industrial district with a specific specialization and integration (e.g. in a productive sector like light mechanical industries) may develop an exchange environment with these features: a few shared ontologies, a coordination structure guiding the development effort, tight integration towards the market. This scenario is captured in our architecture by SINodes. The "wide-shallow" scenario is an open-ended one with much in the way of cooperation but a strong autonomy of the participants, so that no globally shared ontology is available a priori (although it may tend to emerge over time). This scenario is captured in our architecture by means of the brokering and the query agents networks.

3.2 SINodes

SINodes are mediator-based systems, each including a Virtual Data Store, an Ontology Builder, and a Query Manager. In [4, 2] we proposed the mediator-based system *MOMIS* (Mediator envirOnment for Multiple Information Sources) as a group of tools to provide an integrated access to heterogeneous information. More to face the issues related to scalability in the large-scale, in [3] we propose the exploitation of *agents* in the information integration area, and, in particular, their integration in the *MOMIS* infrastructure.

The global ontology of a SINode is an integrated Global Virtual View (GVV) of the managed local ontologies and a mapping between the global integrated view itself and the integrated local ontologies. The GVV is *annotated*: each concept, i.e., each Global Class and each Global Attribute, is associated with a name and set of meanings w.r.t. a to common lexicon ontology (as WordNet/EuroWordnet). One of the main innovation that we propose in the SEWASIE system with respect to MOMIS is a semi-automatic methodology to generate the GVV annotation by starting from annotations of the local sources and mappings between the GVV and the local classes.

In SEWASIE the ODL_{I^3} language [4] was selected to be used to describe heterogeneous schemata of data sources in a common way. In the context of the global ontology of a SINode, ODL_{I^3} introduces constructors useful both in the integration process and in the GVV representation. In particular, intensional relationships expressing inter-schema knowledge for the source schemas defined between classes and attributes names (terms) are supported: SYN (Synonyms), BT (Broader Terms), NT (Narrower Terms) and RT (Related Terms). With reference to the "super–peer network" architecture proposed in [11], where metadata for a small group of peers is centralized onto a single super–peer, we have

that *ontologies* for a small group of peers are *integrated* onto a single super-peer which is an SINodes.

3.3 Brokering Agents

A brokering agent knows about the ontologies which are present in the underlying SINodes, has some information about related (to its own) ontologies in other nodes and has generic information about other ontologies.

The main task of a brokering agent is to provide metadata about the SINodes under its control to the query agents, but each brokering agent is connected to other brokering agents, thereby forming a network of peers where each agent has knowledge about a certain subpart of the network.

The depth of the information of the BA becomes more and more shallow with the distance beetween the ontologies for which it is "expert" (those of its underlying SINodes) and other ontologies covered within the system. Its information on other (non local) ontologies is incomplete.

In the wide-shallow scenario each peer brokering agent is a super-peer and it will develop its own partial view of the world, which is as global as it may be under the circumstances.

Thus the crucial role of the brokering agent in this scenario is the creation and maintenance of the map of semantic relationships among concepts which belong to different SINodes. These relationships are created by the brokering agent which, in a (semi-) automatic way, analyses the "meaning" of the concepts in different ontologies and tries to discover semantic relationships among them by exploiting lexical ontologies such as WordNet/EuroWordNet. Once the repository of these mappings has been created, the brokering agent is in charge of its maintenance: changes in the network have to be integrated to make the repository consistent with the new scenario.

The ontology language within a SINode and the relationships language is ODL_{I^3}. We will analyse possible extensions to ODL_{I^3}, such as constructors expressing other relationships (e.g. antonym), or expressing references to other ontologies, or attributing a likelihood percentage to each relationship, where the percentage value will change depending on the context and the validity of other relationships.

Moreover, mapping functions, which have to support the definition of relationships of concepts of different SINodes, will build upon the concepts of similarity and linguistic meaning. The concept of similarity [14] is the basis for a global strategy towards the identification and description of relationships among

ontologies, and it is currently studied by several research and standardization groups within the Semantic Web initiative.

Let O be the incoming ontology description. Then a typical approach is based on a number of steps namely

- **a normalization step**, that brings O to a common syntax, structure, and language expression,
- **a lexical similarities identification step** (supported by lexical ontologies like WordNet), and
- **a property (attributes, relations) similarity identification step** (it should be noted that if a mapping exists, then it should be established in both directions between the classes involved).

This is a basic strategy, which may be developed in a more articulate fashion with special purpose functions (depending on the goals and strategies of a BA).

3.4 Query Agents

The query agent is the network query manager and "motion item" of the system, and it should be the only carrier of information among the users and the system. Therefore it should be able to do several jobs:

- carrying a query plus the relevant pieces of the user ontology/profile which may help the brokering agents qualify the semantics of the query
- defining the query plan, doing the query rewriting for a specific SINode, and merging the results from several SINodes
- processing the information given by the BA and identifying the SINodes to be accessed for answering the query, and on which further BA to contact to possible get more answer to the query
- carrying back the results (both data and metadata)

4 An example

We show our approach applied on the information provided by four real web sites, two Italians and two Americans, that describe enterprises and products. "Comitato Network Subfornitura" Italian web site (http://www.subfor.net) allows the users to query an online database where detailed information on Italian enterprises and their products are available. The second Italian website (http://www.ingromarket.it) describes products and information about one of the major center for wholesales. The American Apparel Producers' Network

web site (http://www.usawear.org) gives information about products, goods and services related to the textile business. Finally, we analyze a web portal (http://www.fibre2fashion.com) where garment, textile, fashion products are presented. We suppose that the Italian web sites have been integrated into an SINode (SINode$_{IT}$); the obtained GVV (see figure 2) contains three global classes: `Enterprise`, `Manufacturer` and `Category`, where manufacturers are enterprises producing materials belonging to a single category. The American web sites have been integrated into another SINode (SINode$_{US}$); the obtained GVV (see figure 3) contains two global classes: `ManufacturingBusiness` and `ProductClassification`. We extract the ODL$_{I^3}$ schema of these web sites in two steps. On the first step a wrapper translates the HTML pages into XML ones and extract the corresponding DTD files. Then we use our wrapper to translate the XML information in ODL$_{I^3}$ language to integrate the sources in the system.

We tested and reviewed many research and commercial tools, such as Lixto [1], RoadRunner [10], Andes [15], and we select Lixto as the most suitable for our approach. By providing a fully visual and interactive user interface, Lixto assists the user to create wrapper programs in a semi-automatic way. Once the wrapper is built, it can be applied automatically to continually extract relevant information from a permanently changing web page.

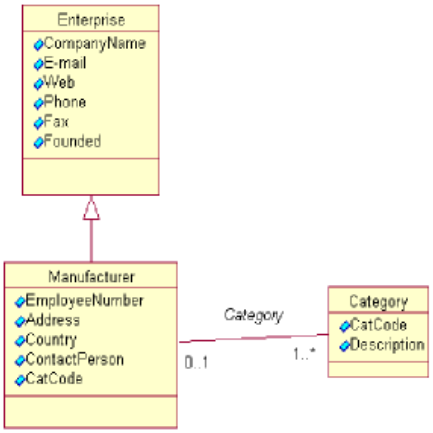

Fig. 2. SINode$_{IT}$ GVV

Let us suppose that these two SINodes have been related to the same BA, as they refer to the same domain ontology. The following semantic relationships:

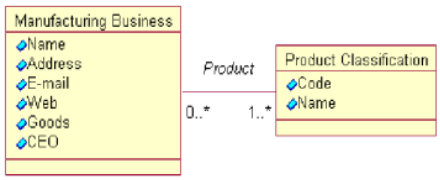

Fig. 3. SINode$_{US}$ GVV

`Manufacturer SYN Manufacturing Business`

`Category SYN Product Classification`

have been discovered by the BA on the basis of the meanings stated in the annotated GVVs and thus are hold by it.

Let us suppose that the user wants to search for information about the specific textile domain. The user interface helps him to write a query and it is connected to a BA; on the basis of this query the SEWASIE system has to individuate the involved further SINodes to generate and solve the executable query. This process is done by means of query agents and by exploiting the information held in the BAs as follows:

1. The user types the query. For example the query is the following: "Search *Manufacturer* of *leather trousers* which has been *founded* before 1999"
2. The user interface recognizes the "keywords" (in italic in the example), disambiguates them looking up the BA ontology, and rewrites the query in terms of simple concepts and send it to a QA. The QA interacts with the user BA that individuates the GVV classes that match with the concepts previously extracted.
3. For each involved SINode the BA produces one/more conjuctive queries and send them to the query agent (e.g. regarding the first SINode: "Select Manufacturer founded before 1999 and category.description is leather trousers" and "Select [Manufacturer Business] founded before 1999 and [Product classification].description is leather trousers").
4. The queries are sent to the QA that is in charge of the generation and execution of the query plans.
5. The query sent to the SINode is rewritten and solved within it.
6. These results of the queries of each node are then fused by the QA and the answer is sent to the user interface.

5 Conclusion

In this paper we provided a description of the EU research project SEWASIE and a proposal of an architectural evolution of the SEWASIE system in the direction of peer-to-peer paradigm. The SEWASIE project has the aim to design and implement an advanced search engine enabling intelligent access to heterogeneous data sources on the web using community-specific multilingual ontologies. The next step of research is the design of the SEWASIE system architecture following the P2P paradigm. This step includes the study and development of a coordination language and of a negotiation protocol among peers.

Acknowledgments

This work is supported in part by the 5th Framework IST programme of the European Community through project SEWASIE within the Semantic Web Action Line. The SEWASIE consortium comprises in addition to the author' organization (Sonia Bergamaschi is the coordinator of the project), the Universities of Aachen RWTH (M. Jarke), Roma La Sapienza (M. Lenzerini, T. Catarci), Bolzano (E. Franconi), as well as IBM Italia, Thinking Networks AG and CNA as user organization.

References

1. R. Baumgartner, S. Flesca, and G. Gottlob. Visual web information extraction with lixto. In *Proceedings of VLDB 2001*, pages 119–128, 2001.
2. D. Beneventano, S. Bergamaschi, S. Castano, A. Corni, R. Guidetti, G. Malvezzi, M. Melchiori, and M. Vincini. Information integration: The momis project demonstration. In *Proceedings of 26^{th} VLDB, September, 2000, Cairo, Egypt*, pages 611–614. Morgan Kaufmann, 2000.
3. S. Bergamaschi, G. Cabri, F. Guerra, L. Leonardi, M. Vincini, and F. Zambonelli. Exploiting agents to support information integration. *International Journal on Cooperative Information Systems*, 11(3), 2002.
4. S. Bergamaschi, S. Castano, D. Beneventano, and M. Vincini. Semantic integration of heterogenous information sources. *Journal of Data and Knowledge Engineering*, 36(3):215–249, 2001.
5. S. Bergamaschi, S. Castano, and M. Vincini. Semantic integration of semistructured and structured data sources. *SIGMOD Records*, 28(1), March 1999.
6. P. A. Bernstein, F. Giunchiglia, A. Kementsietsidis, J. M., L. Serafini, and I. Zaihrayeu. Data management for peer-to-peer computing: A vision. In *Proceedings of Fifth WebDB2002 Madison, Wisconsin*, Jun 2002.
7. S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. 7th Int. World Wide Web Conf.*, 14–18 April 1998.

8. M.J. Carey, L.M. Haas, P.M. Schwarz, M. Arya, W.F. Cody, R. Fagin, M. Flickner, A. Luniewski, W. Niblack, D. Petkovic, J. Thomas II, J.H. Williams, and E.L. Wimmers. Towards heterogeneous multimedia information systems: The garlic approach. In *RIDE-DOM*, pages 124–131, 1985.
9. S. S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. D. Ullman, and J. Widom. The tsimmis project: Integration of heterogeneous information sources. In *Proceedings of the 10th Meeting of the Information Processing Society of Japan, October 1994*, pages 7–18, 1994.
10. V. Crescenzi, G. Mecca, and P. Merialdo. Roadrunner: automatic data extraction from data-intensive web sites. In *Proceedings of 2002 ACM SIGMOD Conference*, 2002.
11. N Daswani, H. Garcia-Molina, and B. Yang. Open problems in data-sharing peer-to-peer systems. In *Proceedings of the 9th ICDT*, 2003.
12. L. Gong. Industry Report: JXTA: A Network Programming Environment. *IEEE Internet Computing*, 5(3):88–95, 2001.
13. A. Y. Levy, A. Rajaraman, and J. J. Ordille. Querying heterogeneous information sources using source descriptions. In *Proc. of VLDB 1996*, pages 251–262, 1996.
14. A. Maedche, B. Motik, N. Silva, and R. Volz. MAFRA – A MApping FRAmework for distributed ontologies. *LNCS*, 2473, 2002.
15. RJussi Myllymaki. Effective web data extraction with standard xml technologies. In *Proceedings of WWW 2001*, pages 689–696, 2001.
16. A. Oram. *Peer-to-Peer: Harnessing the Power of Disruptive Technologies*. O'reilly, 2002.
17. M. T. Ozsu and P. Valduriez, editors. *Principles of Distributed Database Systems*. Prentice Hall, 1999.
18. J.M. Smith, P.A. Bernstein, U. Dayal, N. Goodman, T. Landers, K.W.T. Lin, and E. Wong. Multibase – integrating heterogeneous distributed database systems. In *Proceedings of 1981 National Computer Conference*, pages 487–499. AFIPS Press, 1981.
19. F. van Harmelen, H. Stuckenschmidt, and F. Giunchiglia. Query processing in ontology-based peer-to-peer systems. Technical Report DIT-02-096, University of Trento, 2002.
20. G. Wiederhold. Mediators in the architecture of future information systems. *IEEE Computer*, 25:38–49, 1992.
21. W.Litwin, L. Mark, and N. Roussopoulos. Interoperability of multiple autonomous databases. *ACM Computing Surveys*, 22(3):267–293, 1990.